Week 05 - Exercise 9

Binay Jena

October 3rd, 2020

Student Survey

Problem Statement : As a data science intern with newly learned knowledge in skills in statistical correlation and R programming, you will analyze the results of a survey recently given to college students. You learn that the research question being investigated is: "Is there a significant relationship between the amount of time spent reading and the time spent watching television?" You are also interested if there are other significant relationships that can be discovered? The survey data is located in this StudentSurvey.csv file.

a. Covariance of Survey Variables

```
## Use R to calculate the covariance of the Survey variables and provide an
## explanation of why you would use this calculation and what the results
## indicate.
cov(students_df)
```

```
##              TimeReading       TimeTV  Happiness      Gender
## TimeReading   3.05454545 -20.36363636 -10.350091 -0.08181818
## TimeTV      -20.36363636 174.09090909 114.377273  0.04545455
## Happiness   -10.35009091 114.37727273 185.451422  1.11663636
## Gender       -0.08181818   0.04545455   1.116636  0.27272727
```

Covariance is the simplest way to look or compare the two variables. It helps in understanding whether the two variables in question simultaneously i.e. co-vary with each other. This metric is widely accepted as an indicator of related-ness amongst two variables. A positive covariance indicates directional bias or movement in one of variables from the mean would also mean same directional deviation with the other variable w.r.t the eman. If this deviation is in opposite direction from the mean then the covariance is negative.

Based on the table above, we can say

1. TimeReading and TimeTV vary in opposite ways as the value is -20 approx.
2. TimeReading and Happiness also vary negatively as value is -10 approx
3. TimeReading and Gender also shows -ve relation
4. TimeTV and Happiness are varying +ve as values is 114 approx
5. TimeTV and Gender also shows +ve relation
6. Happiness and Gender also shows +ve relation

b. Examine the Survey data variables. What measurement is being used for the variables? Explain what effect changing the measurement being used for the variables would have on the covariance calculation. Would this be a problem? Explain and provide a better alternative if needed.

It seems, we are having following measurements for each of the survey variables

1. TimeReading: in hours
2. TimeTV: in minutes
3. Happiness: some numeric score could be percentages too
4. Gender: Two values 0 and 1. Each number might represent either Male or Female

Covariance is a measure of relationship. So if we change the measurement the covariance will change. Covariance is a non standardized measurement.

This dependence on scale of measurement is a problem because we cannot compare covariances in an objective way and so we cannot say whether to the covariance is large or small. We may say it objectively, if both datasets have same units(for eg. TimeTV and TimeReading if converted to same unit). However that may not be possible with all variables, as you can see for Happiness and TimeReading or Happiness and TimeTV.

Better alternative method is to standardize this, which is done by using Pearson Correlation Coefficient. This standardized covariance is known as Pearson Coefficient and calculated as below.

$r = cov_{xy} / s_x s_y$

Here, cov_{xy} : Covariance of X and Y, s_x : Standard Deviation of X, s_y : Standard Deviation of Y

c. Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?

We may use the Pearson Correlation Coefficient test to predict. The reason for choosing could be that all variables are intervals except the gender.

And also we may use it, if one of the variables is categorical with two categories, as for our gender variable

d. Perform a correlation analysis of:

If you look at the Pearson Coefficient or Correlation values. Following could be deduced.

1. All variables

```
# default method is Pearson
cor(students_df)
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
## TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
## Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
## Gender      -0.08964215  0.006596673  0.1570118  1.000000000
```

```r
#cor(students_df, method = "spearman")
#cor(students_df, method = "kendall")
```

2. A single correlation between two a pair of the variables

```r
# default is Pearson
cor(students_df$TimeReading, students_df$TimeTV)
```

```
## [1] -0.8830677
```

```r
# This provides more details
cor.test(students_df$TimeReading, students_df$TimeTV)
```

```
##
##  Pearson's product-moment correlation
##
## data:  students_df$TimeReading and students_df$TimeTV
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9694145 -0.6021920
## sample estimates:
##       cor
## -0.8830677
```

3. Repeat your correlation test in step 2 but set the confidence interval at 99%

```r
# This provides more details
cor.test(students_df$TimeReading, students_df$TimeTV, conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  students_df$TimeReading and students_df$TimeTV
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.9801052 -0.4453124
## sample estimates:
##       cor
## -0.8830677
```

4. Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

```r
# Calculating Correlation Coefficient
cor(students_df)
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
## TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
## Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
## Gender      -0.08964215  0.006596673  0.1570118  1.000000000
```

Based on the matrix for correlation and Sams Tips[1] for the full dataframe, we can say the following

1. TimeReading and TimeTV have a large -ve correlation.
2. TimeReading and Happiness have a medium -ve correlation
3. TimeReading and Gender have negligible -ve correlation
4. TimeTV and Happiness have large +ve correlation
5. TimeTV and Gender have negligible + correlation
6. Gender and Happiness have negligible +ve correlation

e. Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.

```r
# Calculating Correlation Coefficient
cor(students_df)
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
## TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
## Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
## Gender      -0.08964215  0.006596673  0.1570118  1.000000000
```

```r
# Calculating coefficent of determination - R^2
cor(students_df)^2
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 1.000000000 0.7798085292 0.18910873 0.0080357143
## TimeTV      0.779808529 1.0000000000 0.40520352 0.0000435161
## Happiness   0.189108726 0.4052035234 1.00000000 0.0246527174
## Gender      0.008035714 0.0000435161 0.02465272 1.0000000000
```

Although we cannot make direct conclusions about causality from a correlation, we can take the correlation coefficient a step further by squaring it. The correlation coefficient squared (known as the coefficient of determination, R^2) is a measure of the amount of variability in one variable that is shared by the other. In our student survey example the correlation coefficient tells us that the watching TV is negatively related to reading. However we don't know how much percent of affected reading time is because of watching TV. This is where **Coefficient of Determination** comes handy. It shows us what percent of reading is affected by watching TV. So above R^2 matrix shows that the 77% of the time the reading is affected by watching TV.

[1]Based on Sams Tips(Discovering Statitics Using R)1. +- 0.1 - means small effect 2. +- 0.3 tp 0.5 - medium effect 3. over +- 0.5 large effect

f. Based on your analysis can you say that watching more TV caused students to read less? Explain.

Yes, Watching TV causes Students to read less. Based on correlation test of student survey attributes we can say reading is affected by watching TV. Also as we have seen coefficient of determination also shows as much as 77% of the time reading time is affected by watching TV.

g. Pick three variables and perform a partial correlation, documenting which variable you are "controlling". Explain how this changes your interpretation and explanation of the results.

Lets do the partial correlation by controlling the gender variable.

```
library(ggm)
# Partial Correlation, controlling variable is Gender
pcor(c("TimeReading","TimeTV", "Gender"), var(students_df))
```

```
## [1] -0.8860628
```

```
# coefficient of determination - R^2
pcor(c("TimeReading","TimeTV", "Gender"), var(students_df))^2
```

```
## [1] 0.7851073
```

Partial correlation analysis using TimeTv, TimeReading and Happiness shows that the time watching TV is negatively affecting reading time. Also when we keep Happiness constant doesn't affect much the relation between watching TV and reading time. With correlation test we had r = -0.88 where as with partial test we get partial correlation of -0.87.